TRABALHO TECH CHALLENGE – FASE 3

Desenvolvimento de um protótipo de assistente virtual médico, usando dados fictícios de pacientes e bases médicas de conhecimento.

MARCELO FERNANDES
6IADT - RM 366035

OBJETIVOS DO TRABALHO

- ❑ Desenvolvimento de um assistente virtual médico
- ❑ Refinamento (fine-tuning) de uma LLM a partir de dados médicos internos
- ❑ Pré-processamento dos dados com foco em limpeza, anonimização e curadoria
- ❑ Estabelecer limites de atuação (guardrails) do assistente virtual para evitar sugestões impróprias
- ❑ Garantir explicabilidade das respostas, indicando fontes de informação usadas como base para elaboração da resposta
- ❑ Organizar todas as sáidas em um repositório no Github

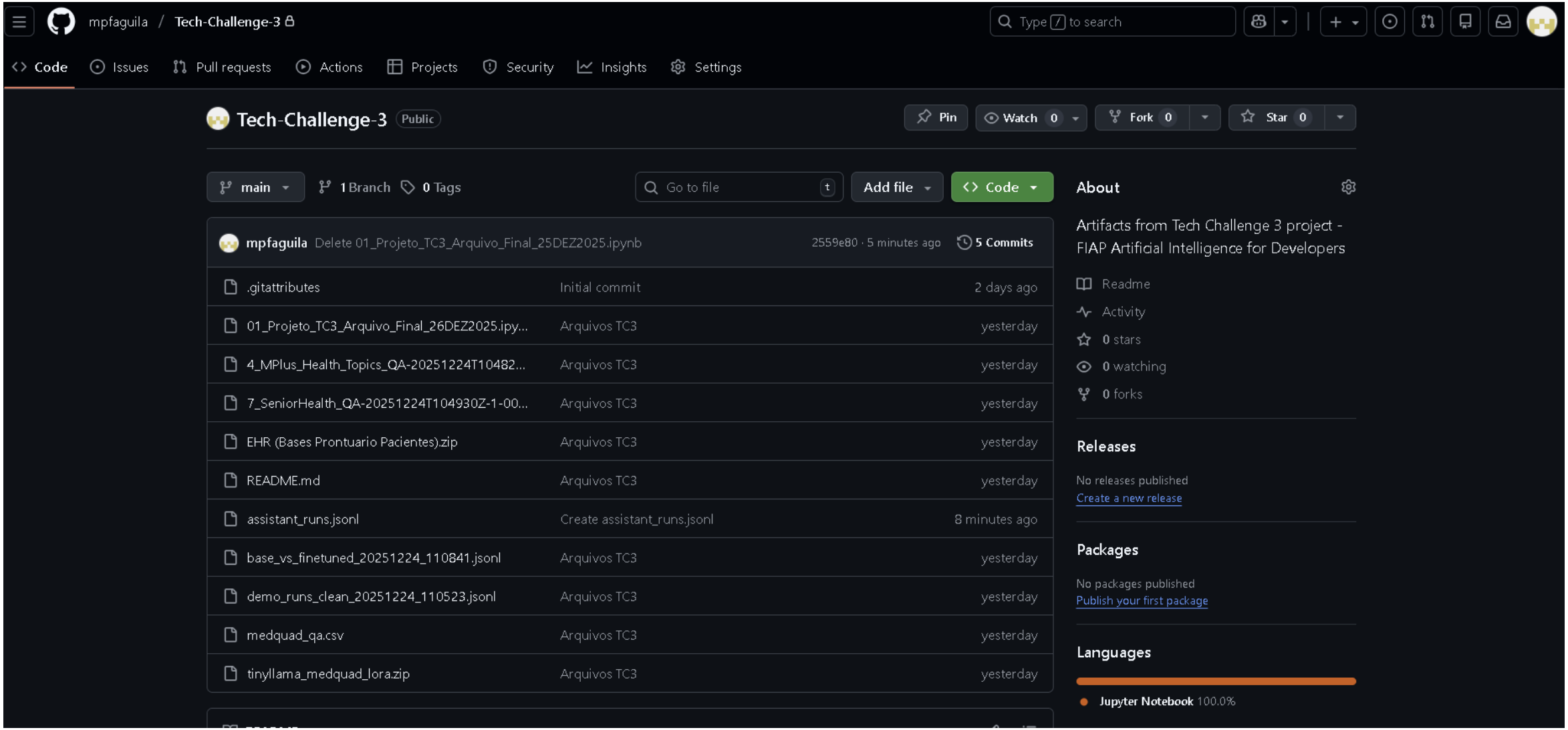
AMBIENTE UTILIZADO PARA O TRABALHO

- Google Colab Pro +, na plataforma Google Cloud
- Uso de máquinas GPU L4, com maior poder de processamento
- Notebook único, com todas as etapas do trabalho, 100% em linguagem Python
- Do dataset sugerido na descrição do Tech Challenge 3, foram utilizados dados do MedQuad e dados específicos foram filtrados, focando em assuntos de saúde geral (**4_MPlus_Health_Topics_QA**), bem como dados associados a pacientes da 3a idade (**7_SeniorHealth_QA**). Como esses dados estão todos em inglês, toda a estrtuturação do assistente virtual medico foi feita no idioma Inglês, sem tradução para o Português.

OUTPUTS GERADOS DURANTE O TRABALHO

- ❑ CSVs do EHR (prontuário) sintético
- ❑ Índice vetorial FAISS + metadados
- ❑ Adapter LoRA do modelo fine-tuned
- ❑ Logs estruturados das execuções
- ❑ Notebook Python com todo o processo dividido em blocos,
- ❑ README com as instruções
- ❑ Relatório técnico
- ❑ Vídeo do trabalho

REPOSITÓRIO GITHUB

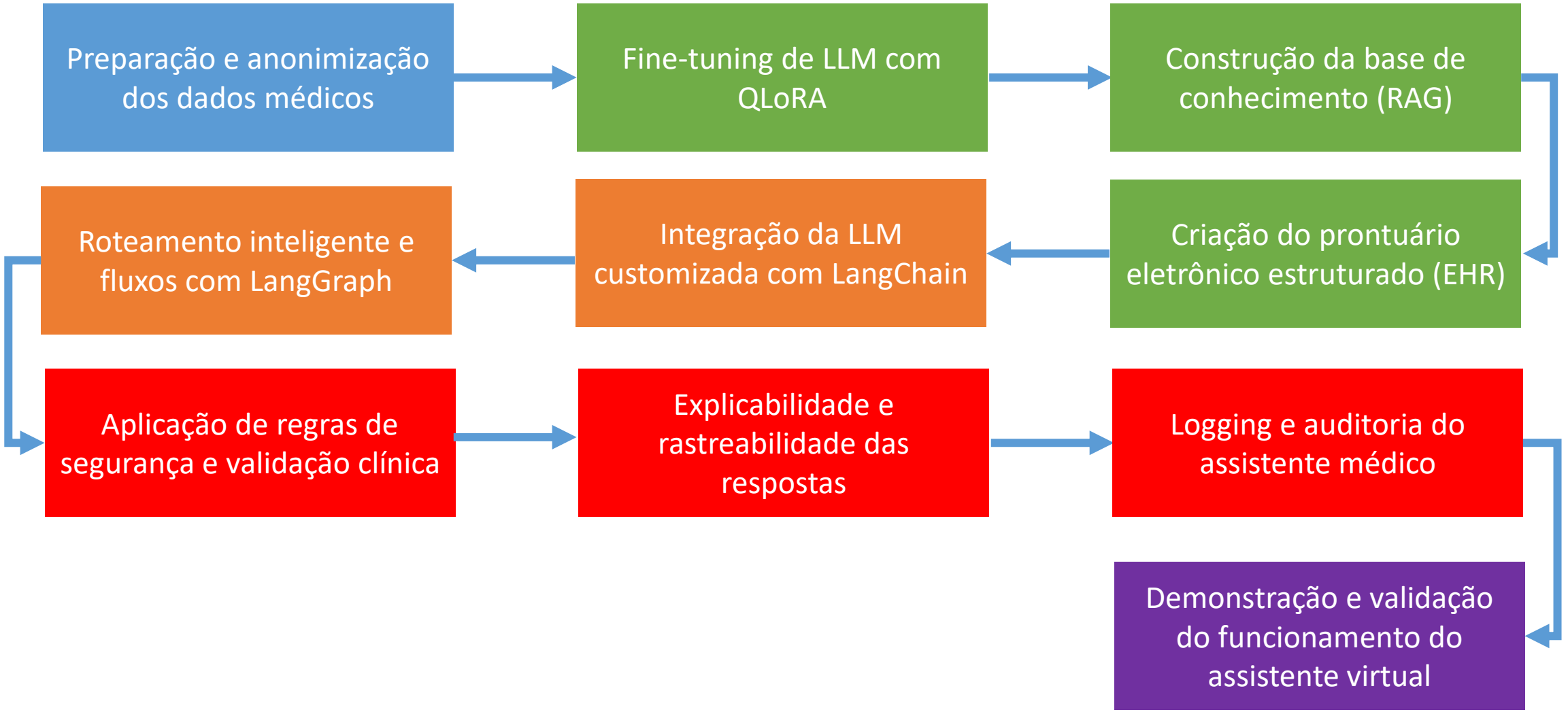


Fonte: [mpfaguila/Tech-Challenge-3: Artifacts from Tech Challenge 3 project - FIAP Artificial Intelligence for Developers]

ETAPAS DO DESENVOLVIMENTO DO ASSISTENTE VIRTUAL MÉDICO

Fluxo de trabalho utilizado

AS 10 ETAPAS DO DESENVOLVIMENTO DO ASSISTENTE VIRTUAL



Dados e Treinamento

RAG e EHR (prontuário do paciente)

LangChain e LangGraph

Segurança e Auditoria

Demo e Validação

ETAPA 1 – PREPARAÇÃO E ANONIMIZAÇÃO DOS DADOS MÉDICOS

PREPARAÇÃO E ANONIMIZAÇÃO DOS DADOS MÉDICOS

Seleção de datasets clínicos públicos

- ❑ Utilização do ***MedQuAD** como *proxy* de conhecimento médico institucional
- ❑ Conjunto composto por perguntas e respostas clínicas revisadas por especialistas

Criação de dados sintéticos de prontuário eletrônico (EHR)

- ❑ Geração artificial de pacientes, visitas, diagnósticos (CID-10), exames laboratoriais e prescrições
- ❑ Estrutura inspirada em sistemas de prontuário minimamente reais, sem uso de dados sensíveis reais

Anonimização e conformidade com privacidade

- ❑ Ausência total de dados pessoais reais (nomes, documentos, contatos)
- ❑ Uso exclusivo de identificadores sintéticos (IDs numéricos)
- ❑ Adequação conceitual à LGPD (Lei Geral de Proteção a Dados e boas práticas de dados em saúde

Pré-processamento e padronização dos dados

- ❑ Limpeza de campos textuais e normalização de formatos
- ❑ Conversão para estruturas tabulares (CSV) e textuais (QA)
- ❑ Organização consistente para consumo por pipelines de ML e LangChain

Curadoria e separação por finalidade

- ❑ Dados clínicos textuais → treinamento e RAG da LLM
- ❑ Dados estruturados (EHR) → consultas determinísticas e factualidade
- ❑ Redução de risco de alucinação ao separar conhecimento de fatos do paciente

***Fonte:** (abachaa/MedQuAD: Medical Question Answering Dataset of 47,457 QA pairs created from 12 NIH websites)

ETAPA 2 – FINE TUNING DE LLM COM QLORA

FINE TUNING DE LLM COM QLORA (1/2)

Seleção do modelo base de linguagem (LLM)

- ❑ Utilização de um modelo open-source leve (TinyLlama – 1,1 bilhão de parâmetros)
- ❑ Escolha orientada a viabilidade computacional em ambiente Google Colab

Definição da estratégia de fine-tuning eficiente

- ❑ Aplicação da técnica QLoRA (Quantized Low-Rank Adaptation)
- ❑ Redução de custo computacional e memória
- ❑ Ajuste fino sem re-treinamento completo do modelo

Preparação da base para treinamento supervisionado

- ❑ Uso do MedQuAD como proxy de conhecimento médico institucional
- ❑ Estruturação no formato pergunta–resposta (instruction tuning)
- ❑ Garantia de coerência clínica e qualidade textual

Configuração do pipeline de treinamento

- ❑ Quantização do modelo base para 4 bits
- ❑ Definição de hiperparâmetros (batch size, learning rate, epochs)
- ❑ Uso de bibliotecas Hugging Face + PEFT

Execução do fine-tuning supervisionado

- ❑ Treinamento do adapter LoRA sobre o modelo quantizado
- ❑ Aprendizado focado em linguagem e raciocínio clínico
- ❑ Monitoramento do processo de treino

***Fonte:** (abachaa/MedQuAD: Medical Question Answering Dataset of 47,457 QA pairs created from 12 NIH websites)

FINE TUNING DE LLM COM QLORA (2/2)

Validação qualitativa do modelo ajustado

- ❑ Comparação entre respostas do modelo base e do modelo fine-tuned
- ❑ Verificação de maior precisão e contextualização clínica
- ❑ Avaliação manual orientada a casos médicos

***Fonte:** (abachaa/MedQuAD: Medical Question Answering Dataset of 47,457 QA pairs created from 12 NIH websites)

ETAPA 3 - CONSTRUÇÃO DA BASE DE CONHECIMENTO

CONSTRUÇÃO DA BASE DE CONHECIMENTO (1/2)

Definição da estratégia de Recuperação Aumentada por Geração (RAG)

- ❑ Separação entre conhecimento médico textual e dados factuais do paciente
- ❑ Redução de alucinações e aumento da confiabilidade das respostas

Pré-processamento dos dados clínicos

- ❑ Utilização do dataset MedQuAD como base de conhecimento médico
- ❑ Organização dos documentos em formato adequado para indexação
- ❑ Associação de metadados para rastreabilidade das fontes

Geração de embeddings semânticos

- ❑ Conversão dos textos clínicos em vetores numéricos
- ❑ Captura de similaridade semântica entre perguntas e documentos
- ❑ Preparação dos dados para busca vetorial eficiente

Indexação em base vetorial FAISS

- ❑ Criação de índice vetorial para recuperação rápida de contexto
- ❑ Armazenamento local dos embeddings
- ❑ Suporte a consultas semânticas em tempo de execução

***Fonte:** (abachaa/MedQuAD: Medical Question Answering Dataset of 47,457 QA pairs created from 12 NIH websites)

CONSTRUÇÃO DA BASE DE CONHECIMENTO (2/2)

Implementação do mecanismo de recuperação de contexto

- ❑ Busca dos documentos mais relevantes para cada pergunta
- ❑ Seleção dos trechos clínicos mais pertinentes
- ❑ Enriquecimento do prompt da LLM com contexto recuperado

Integração da base RAG ao assistente médico

- ❑ Uso da base vetorial como ferramenta (Tool) no LangChain
- ❑ Disponibilização do conhecimento médico para o fluxo decisório
- ❑ Garantia de respostas fundamentadas em fontes documentadas

Validação das respostas baseadas em RAG

- ❑ Verificação manual da coerência clínica das respostas
- ❑ Checagem da correspondência entre resposta e documentos recuperados
- ❑ Inclusão explícita das fontes utilizadas na resposta final

ETAPA 4 – CRIAÇÃO DO PRONTUÁRIO ELETRÔNICO ESTRUTURADO (EHR)

CONSTRUÇÃO DO PRONTUÁRIO ELETRÔNICO ESTRUTURADO

Modelagem do prontuário eletrônico estruturado (EHR)

- ❑ Definição das entidades principais: paciente, visitas, diagnósticos, exames laboratoriais e prescrições
- ❑ Estrutura inspirada em prontuários reais, porém com dados sintéticos

Geração de dados sintéticos e anonimizados

- ❑ Criação de um conjunto de pacientes e histórico clínico sem dados pessoais reais
- ❑ Uso de IDs numéricos e informações fictícias (conformidade conceitual com LGPD)

Organização em dados tabulares para consulta

- ❑ Separação por domínio (ex.: pacientes.csv, visitas.csv, diagnósticos.csv, labs.csv, prescrições.csv)
- ❑ Padronização de colunas, datas e formatos para facilitar ingestão e auditoria

Implementação de consultas determinísticas ao EHR

- ❑ Funções para extrair resumo do paciente, diagnósticos ativos/crônicos, visitas recentes e exames recentes
- ❑ Consulta estruturada via Pandas, garantindo factualidade dos dados do paciente

Detecção de pendências e gatilhos clínicos

- ❑ Rotinas para identificar exames pendentes, lacunas de acompanhamento e alertas básicos
- ❑ Preparação do EHR para ser usado em fluxos de decisão (LangGraph)

Contextualização das respostas com dados atualizados do paciente

- ❑ Uso do EHR como fonte única para fatos clínicos do paciente
- ❑ Separação entre Fatos do paciente (EHR) vs Conhecimento médico (RAG/LLM)
- ❑ Redução de risco de alucinação e aumento de confiabilidade

ETAPA 5 – INTEGRAÇÃO DA LLM CUSTOMIZADA COM LANGCHAIN

INTEGRAÇÃO DA LLM CUSTOMIZADA COM LANGCHAIN

Definição da arquitetura “LLM + Ferramentas (Tools)”

- ❑ Separação de responsabilidades:EHR Tool → fatos do paciente (determinístico) e RAG Tool → conhecimento médico (documentos)
- ❑ Base para respostas mais seguras e auditáveis

Integração do modelo LLM customizado

- ❑ Uso do modelo ajustado como motor principal do assistente
- ❑ Respostas condicionadas ao contexto vindo das Tools
- ❑ Preparação para operação em pipeline (chain)

Construção do pipeline de resposta com LangChain

- ❑ Entrada do usuário → seleção de ferramentas → agregação de contexto
- ❑ Montagem do prompt final com:
  - ❑ dados do paciente (EHR)
  - ❑ documentos recuperados (RAG)
  - ❑ regras de segurança (guardrails)
- ❑ Geração de resposta final mais contextualizada

Padronização de saída com fontes e rastreabilidade

- ❑ Indicação explícita se a resposta usou: EHR, RAG ou ambos
- ❑ Inclusão de “fontes” com origem determinística (evita invenção)
- ❑ Preparação para logging e auditoria

ETAPA 6 – ROTEAMENTO INTELIGENTE E FLUXOS COM LANGGRAPH

ROTEAMENTO INTELIGENTE E FLUXOS COM LANGGRAPH (1/2)

Definição da estratégia de roteamento inteligente

- Classificação automática das perguntas médicas em três categorias:
  - EHR → perguntas sobre dados do paciente
  - RAG → perguntas sobre conhecimento médico geral
  - AMBOS → perguntas que exigem dados do paciente + conhecimento clínico

Implementação do nó de decisão (roteador)

- Uso da LLM para interpretar a intenção da pergunta
- Determinação explícita das ferramentas que devem ser acionadas
- Evita consultas desnecessárias e melhora eficiência

Construção do grafo de execução com LangGraph

- Definição de nós especializados:
  - nó EHR
  - nó RAG
  - nó AMBOS
- Modelagem do fluxo como um grafo de estados

Execução condicional das ferramentas

- Ativação apenas das Tools necessárias conforme a rota escolhida
- Garantia de acesso controlado aos dados clínicos
- Separação clara entre etapas determinísticas e geração textual

ROTEAMENTO INTELIGENTE E FLUXOS COM LANGGRAPH (2/2)

Agregação do contexto para resposta final

- ❑ Consolidação dos resultados vindos do EHR e/ou RAG
- ❑ Preparação do prompt final para a LLM
- ❑ Manutenção da rastreabilidade das informações utilizadas

Controle e previsibilidade do fluxo clínico

- ❑ Fluxo explícito e auditável de decisões
- ❑ Facilidade para validação, testes e extensão futura
- ❑ Redução de comportamentos não determinísticos do assistente

ETAPA 7 – APLICAÇÃO DE REGRAS DE SEGURANÇA E VALIDAÇÃO CLÍNICA

APLICAÇÃO DE REGRAS DE SEGURANÇA E VALIDAÇÃO CLÍNICA (1/2)

Definição clara dos limites de atuação do assistente

- ❑ O assistente atua apenas como apoio à decisão clínica
- ❑ Proibição explícita de diagnósticos definitivos ou prescrições médicas
- ❑ Exigência de validação humana para qualquer conduta clínica

Aplicação de guardrails clínicos no prompt da LLM

- ❑ Regras explícitas para evitar recomendações impróprias
- ❑ Orientação para respostas informativas e explicativas
- ❑ Controle do escopo de atuação do modelo

Separação entre fatos clínicos e conhecimento médico

- ❑ Fatos do paciente obtidos exclusivamente via EHR estruturado
- ❑ Conhecimento médico obtido via RAG
- ❑ Evita inferências ou suposições não verificáveis

Validação no uso correto das ferramentas

- ❑ Restrições sobre quando e como o EHR pode ser consultado
- ❑ Uso condicionado das ferramentas conforme o fluxo do LangGraph
- ❑ Redução de acesso indevido ou desnecessário a dados clínicos

APLICAÇÃO DE REGRAS DE SEGURANÇA E VALIDAÇÃO CLÍNICA (2/2)

Controle de interpretação de exames e dados sensíveis

- ❑ Evita classificações automáticas (ex.: “alto”, “baixo”, “normal”) sem contexto clínico completo
- ❑ Apresentação dos valores de forma descritiva e contextualizada
- ❑ Redução de risco de conclusões clínicas incorretas

Explicabilidade e transparência das respostas

- ❑ Indicação explícita das fontes utilizadas (EHR e/ou RAG)
- ❑ Justificativa textual do raciocínio seguido pelo assistente
- ❑ Aumento da confiança e auditabilidade das respostas

ETAPA 8 – EXPLICABILIDADE E TRANSPARÊNCIA DAS RESPOSTAS

EXPLICABILIDADE E TRANSPARÊNCIA DAS RESPOSTAS

Indicação explícita das fontes utilizadas na resposta

- ❑ Identificação clara se a resposta utilizou:
  - ❑ Dados do prontuário eletrônico (EHR)
  - ❑ Base de conhecimento médico (RAG)
  - ❑ Ou ambos (EHR + RAG)
- ❑ Evita respostas sem fundamentação verificável

Separação clara entre fatos e conhecimento médico

- ❑ Fatos clínicos apresentados exclusivamente com base no EHR
- ❑ Conhecimento médico contextual fornecido a partir de documentos recuperados
- ❑ Redução de inferências e alucinações

Justificativa do raciocínio seguido pelo assistente

- ❑ Explicação textual de como a resposta foi construída
- ❑ Indicação das informações relevantes consideradas no processo
- ❑ Aumento da confiança do usuário médico

Padronização da saída do assistente

- ❑ Estrutura de resposta consistente (resposta + fontes)
- ❑ Evita variações imprevisíveis na apresentação das informações
- ❑ Facilita validação clínica e revisão humana

ETAPA 9 – LOGGING E AUDITORIA DO ASSISTENTE MÉDICO

LOGGING E AUDITORIA DO ASSISTENTE MÉDICO

Logging estruturado das interações

- ❑ Registro sistemático de cada execução do assistente
- ❑ Armazenamento de entradas, decisões e saídas em formato estruturado (JSON)

Registro do fluxo de decisão do assistente

- ❑ Armazenamento da rota escolhida (EHR, RAG ou BOTH)
- ❑ Identificação das ferramentas acionadas em cada interação
- ❑ Transparência sobre o comportamento do sistema

Rastreabilidade das informações utilizadas

- ❑ Registro das fontes consultadas (EHR e/ou RAG)
- ❑ Associação das respostas aos dados e documentos utilizados
- ❑ Suporte à auditoria clínica e técnica

Suporte à validação humana e compliance

- ❑ Possibilidade de revisão posterior das respostas geradas
- ❑ Apoio à conformidade com boas práticas e requisitos regulatórios
- ❑ Facilita investigação de erros ou inconsistências

Base para monitoramento e melhoria contínua

- ❑ Análise de padrões de uso do assistente
- ❑ Identificação de falhas, ambiguidades ou riscos
- ❑ Insumos para ajustes de prompts, regras e modelos

ETAPA 10 – DEMONSTRAÇÃO E VALIDAÇÃO DO FUNCIONAMENTO DO ASSISTENTE VIRTUAL

DEMONSTRAÇÃO E VALIDAÇÃO DO FUNCIONAMENTO

Execução de cenários clínicos simulados

- ❑ Demonstração do assistente respondendo a perguntas clínicas reais
- ❑ Uso de dados de pacientes sintéticos para contextualização
- ❑ Simulação de situações comuns do ambiente hospitalar

Comparação entre o modelo base e o modelo refinado

- ❑ Apresentação de respostas antes e depois do fine-tuning
- ❑ Avaliação qualitativa da melhoria em precisão e contextualização
- ❑ Justificativa do impacto do ajuste do modelo

Validação de segurança e das regras clínicas

- ❑ Demonstração de respostas respeitando os limites de atuação
- ❑ Evidência da não prescrição e exigência de validação humana
- ❑ Comprovação do funcionamento dos guardrails clínicos

Exibição de explicabilidade, fontes e logs

- ❑ Visualização das fontes utilizadas (EHR e/ou RAG)
- ❑ Demonstração do logging estruturado das interações
- ❑ Prova de rastreabilidade e auditabilidade do sistema

CONCLUSÕES E SUGESTÕES DE MELHORIAS

CONCLUSÕES SOBRE O TRABALHO DE CRIAÇÃO DO ASSISTENTE

- É viável construir um assistente médico seguro utilizando LLMs open-source. Mesmo com recursos computacionais limitados, técnicas como QLoRA permitem adaptação eficiente ao domínio clínico.
- A separação entre dados estruturados (EHR) e conhecimento médico (RAG) é fundamental, já que ajuda a garantir factualidade, reduz alucinações e aumenta a confiabilidade das respostas clínicas.
- O uso de LangChain e LangGraph permite fluxos clínicos controlados e auditáveis. A modelagem em grafo oferece previsibilidade, rastreabilidade e segurança no processo decisório.
- O processo de fine-tuning melhora a qualidade das respostas em contexto médico. O modelo ajustado apresentou respostas mais precisas, contextualizadas e alinhadas à terminologia clínica médica.
- Guardrails clínicos são indispensáveis em aplicações de saúde, pois limites claros de atuação evitam prescrições indevidas e reforçam o papel do médico humano.
- Explicabilidade e transparência aumentam a confiança no sistema. A indicação explícita de fontes e o detalhamento do raciocínio tornam o assistente auditável e confiável.
- Logging estruturado viabiliza governança e melhoria contínua. O registro completo das interações permite auditoria, validação humana e evolução do sistema.
- Dados sintéticos são uma alternativa viável para desenvolvimento e testes em saúde, pois permitem simular cenários clínicos reais respeitando privacidade e LGPD.

POTENCIAIS MELHORIAS

- Integração com prontuários eletrônicos reais (EHRs hospitalares)
  - Substituição do EHR sintético por integração com sistemas reais (FHIR/HL7), mantendo os mesmos mecanismos de segurança e validação
  - Permitiria uso em ambiente clínico real, respeitando políticas de acesso e privacidade

- Interface web para que o assistente médico possa ser facilmente usado por profissionais de saúde, sem necessidade de visualização do código fonte.

- Nesse assistente foram utilizados dados de fontes estrangeiras, no idioma inglês. Uma futura versão poderia trazer dados da realidade local, com informações 100% em Português.